import pandas as pd
import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from google.colab import files
uploaded = files.upload()

Choose Files House Price India.csv

House Price India.csv(text/csv) - 1524561 bytes, last modified: 9/26/2023 - 100% done Saving House Price India.csv to House Price India.csv

import io
df = pd.read_csv(io.BytesIO(uploaded['House Price India.csv']))

df.head()

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	 Buil Yea
0	6762810145	42491	5	2.50	3650	9050	2.0	0	4	5	 192
1	6762810635	42491	4	2.50	2920	4000	1.5	0	0	5	 190
2	6762810998	42491	5	2.75	2910	9480	1.5	0	0	3	 193
3	6762812605	42491	4	2.50	3310	42998	2.0	0	0	3	 200
4	6762812919	42491	3	2.00	2710	4500	1.5	0	0	4	 192

5 rows × 23 columns

df.tail()

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	•••
14615	6762830250	42734	2	1.5	1556	20000	1.0	0	0	4	
14616	6762830339	42734	3	2.0	1680	7000	1.5	0	0	4	
14617	6762830618	42734	2	1.0	1070	6120	1.0	0	0	3	
14618	6762830709	42734	4	1.0	1030	6621	1.0	0	0	4	
14619	6762831463	42734	3	1.0	900	4770	1.0	0	0	3	

5 rows × 23 columns

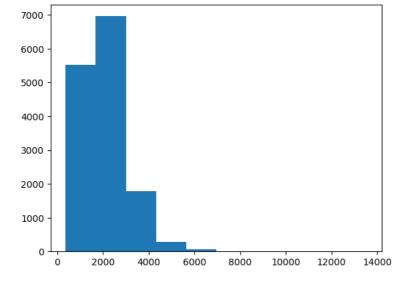
df

number

```
waterfront number condition
                                                                          number
                                             number of living
                       id
                            Date
df.columns
     {\tt Index(['id', 'Date', 'number \ of \ bedrooms', 'number \ of \ bathrooms',}\\
             'living area', 'lot area', 'number of floors', 'waterfront present', 'number of views', 'condition of the house', 'grade of the house',
             'Area of the house(excluding basement)', 'Area of the basement', 'Built Year', 'Renovation Year', 'Postal Code', 'Lattitude', 'Longitude', 'living_area_renov', 'lot_area_renov',
             'Number of schools nearby', 'Distance from the airport', 'Price'],
            dtype='object')
df.dtypes
     id
                                                     int64
     Date
                                                     int64
                                                     int64
     number of bedrooms
     number of bathrooms
                                                   float64
     living area
                                                     int64
     lot area
                                                     int64
     number of floors
                                                   float64
     waterfront present
                                                     int64
     number of views
                                                     int64
     condition of the house
                                                     int64
     grade of the house
                                                     int64
     Area of the house(excluding basement)
                                                     int64
     Area of the basement
                                                     int64
                                                     int64
     Built Year
                                                     int64
     Renovation Year
                                                     int64
     Postal Code
     Lattitude
                                                   float64
     Longitude
                                                   float64
     living_area_renov
                                                     int64
     lot_area_renov
                                                     int64
     Number of schools nearby
                                                     int64
     Distance from the airport
                                                     int64
                                                     int64
     dtype: object
df.info()
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 14620 entries, 0 to 14619
     Data columns (total 23 columns):
      #
          Column
                                                      Non-Null Count Dtype
      0
           id
                                                      14620 non-null
          Date
                                                      14620 non-null int64
      1
           number of bedrooms
                                                      14620 non-null
      2
                                                                        int64
           number of bathrooms
                                                      14620 non-null float64
      3
                                                      14620 non-null
      4
                                                                        int64
           living area
                                                      14620 non-null int64
          lot area
           \quad \text{number of floors} \\
      6
                                                      14620 non-null
                                                                        float64
           waterfront present
                                                      14620 non-null
                                                                        int64
      8
           number of views
                                                      14620 non-null
      9
           condition of the house
                                                      14620 non-null
          grade of the house
                                                      14620 non-null int64
           Area of the house(excluding basement) 14620 non-null
          Area of the basement
                                                      14620 non-null
      12
                                                                        int64
           Built Year
                                                      14620 non-null
                                                                        int64
      13
          Renovation Year
                                                      14620 non-null int64
      14
                                                      14620 non-null
      15
          Postal Code
                                                                        int64
      16
          Lattitude
                                                      14620 non-null
                                                                        float64
      17
          Longitude
                                                      14620 non-null
                                                                        float64
      18
          living_area_renov
                                                      14620 non-null
                                                                        int64
      19
          lot_area_renov
                                                      14620 non-null
                                                                        int64
      20
          Number of schools nearby
                                                      14620 non-null
      21
          Distance from the airport
                                                      14620 non-null
                                                                        int64
                                                      14620 non-null int64
          Price
     dtypes: float64(4), int64(19)
     memory usage: 2.6 MB
df.shape
     (14620, 23)
print(df.describe())
             6.237575e+03
                                67.347991
                                                       0.938719
                                                                              0.769934
     std
```

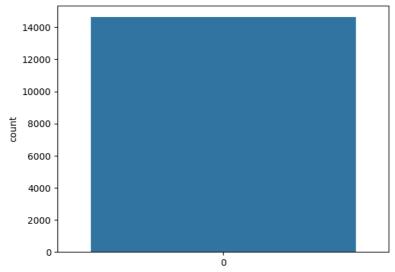
```
D./DZ8Z0E+09
                     42002.000000
                                               4.000000
                                                                     2.500000
15%
       6.762832e+09 42734.000000
                                              33.000000
                                                                     8.000000
max
                          lot area number of floors waterfront present
        living area
                                                             14620.000000
count
       14620,000000
                     1,462000e+04
                                        14620.000000
mean
        2098.262996
                     1.509328e+04
                                            1.502360
                                                                 0.007661
std
         928.275721
                     3.791962e+04
                                            0.540239
                                                                 0.087193
         370.000000
                     5.200000e+02
                                             1.000000
                                                                 0.000000
min
25%
        1440.000000
                     5.010750e+03
                                             1.000000
                                                                 0.000000
50%
        1930.000000
                     7.620000e+03
                                             1.500000
                                                                 0.000000
75%
        2570.000000 1.080000e+04
                                             2.000000
                                                                 0.000000
       13540.000000 1.074218e+06
                                             3.500000
                                                                 1.000000
max
       number of views condition of the house
                                                         Built Year
          14620.000000
                                   14620.000000
                                                       14620.000000
count
                                       3.430506
                                                        1970.926402
mean
              0.233105
std
              0.766259
                                       0.664151
                                                  . . .
                                                          29.493625
                                       1.000000
min
              0.000000
                                                        1900.000000
                                                  . . .
25%
              0.000000
                                        3.000000
                                                        1951.000000
                                                  . . .
50%
              0.000000
                                       3.000000
                                                        1975.000000
                                                 . . .
75%
               0.000000
                                       4.000000
                                                        1997.000000
                                                 . . .
              4.000000
                                       5.000000
                                                        2015.000000
max
       Renovation Year
                           Postal Code
                                           Lattitude
                                                          Longitude
          14620.000000
                          14620.000000 14620.000000
                                                       14620.000000
count
             90.924008 122033.062244
                                           52.792848
                                                        -114.404007
mean
std
            416.216661
                             19.082418
                                            0.137522
                                                           0.141326
min
              0.000000
                         122003.000000
                                            52.385900
                                                        -114.709000
25%
              0.000000
                        122017.000000
                                            52.707600
                                                        -114.519000
50%
              0.000000
                         122032.000000
                                            52.806400
                                                        -114.421000
                        122048.000000
                                            52.908900
75%
              0.000000
                                                        -114.315000
           2015.000000
                                            53.007600
                        122072.000000
                                                        -113.505000
max
                                           Number of schools nearby
       living area renov lot area renov
            14620.000000
                             14620.000000
                                                        14620.000000
count
mean
             1996,702257
                             12753,500068
                                                            2.012244
std
              691.093366
                             26058.414467
                                                            0.817284
                               651.000000
              460.000000
                                                            1.000000
min
25%
             1490.000000
                              5097.750000
                                                            1.000000
50%
             1850.000000
                              7620.000000
                                                            2.000000
75%
             2380.000000
                             10125.000000
                                                            3.000000
max
             6110.000000
                            560617.000000
                                                            3.000000
       Distance from the airport
                                          Price
                    14620.000000
                                   1.462000e+04
count
                        64.950958
                                   5.389322e+05
mean
                        8.936008
std
                                   3.675324e+05
                        50.000000
                                   7.8000000+04
min
                        57.000000
25%
                                   3.200000e+05
50%
                        65.000000
                                   4.500000e+05
75%
                        73.000000
                                   6.450000e+05
max
                        80.000000
                                   7.700000e+06
[8 rows x 23 columns]
```

plt.hist(df['living area'])



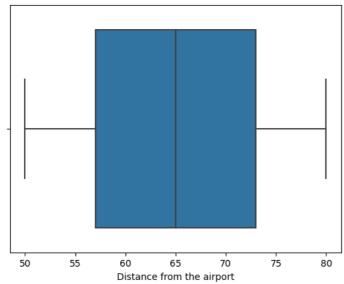
sns.countplot(df['condition of the house'])



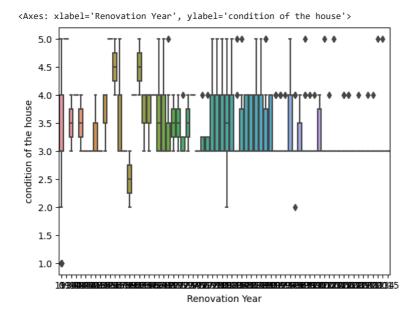


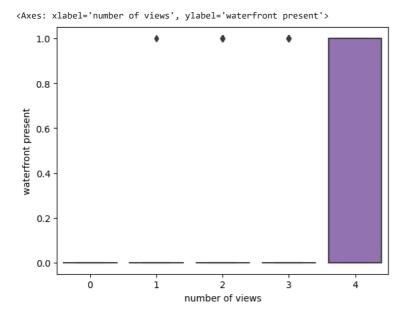
sns.boxplot(x=df['Distance from the airport'])

<Axes: xlabel='Distance from the airport'>

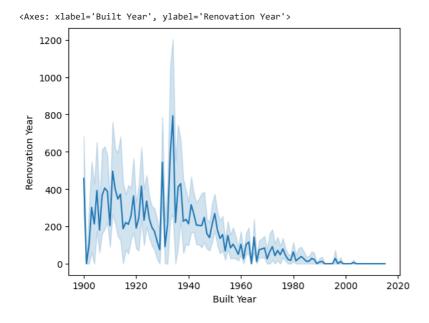


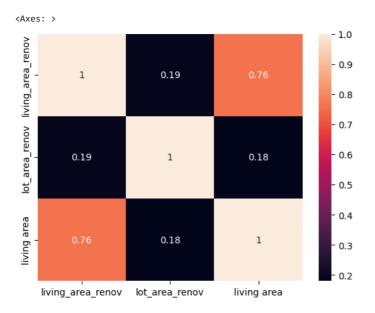
sns.boxplot(x=df['Renovation Year'],y=df['condition of the house'])



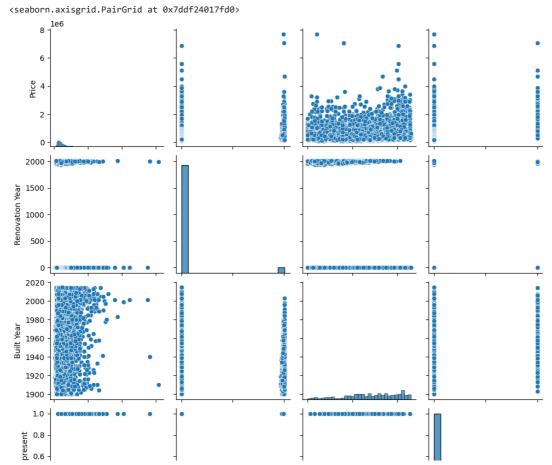


sns.lineplot(x=df['Built Year'],y=df['Renovation Year'])

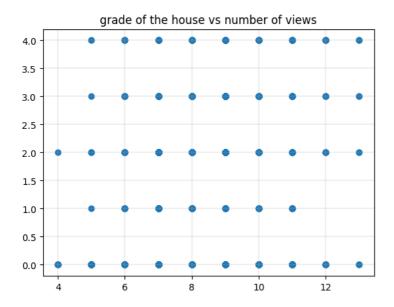




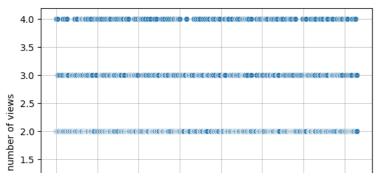
sns.pairplot(df[['Price','Renovation Year','Built Year','waterfront present']])



plt.scatter(df['grade of the house'],df['number of views'])
plt.title("grade of the house vs number of views")
plt.grid(linestyle='-', linewidth=0.3)



sns.scatterplot(df['number of views'])
plt.grid(linestyle='-',linewidth=0.5)



plt.subplots(figsize=(25,25))
sns.heatmap(df.drop(['id'],axis=1).corr(),linewidth=0.4,annot=True)
plt.show()

			_	_			_		_	_	_		_	_	_	_	_	_	_	_	_	_
Date	- 1	-0.016	-0.026	-0.022	0.0044	-0.01	0.012	-0.0048	-0.027	-0.033	-0.016	-0.016	-0.0059	-0.012	0.018	-0.023		-0.032	-5e-05	-0.0041	0.011	-0.028
number of bedrooms	-0.016	1	0.51				-0.0063								-0.044	-0.013			0.029	0.0034	-0.0062	
number of bathrooms	-0.026		1	0.75	0.081		0.06	0.18	-0.13			0.29		0.05	-0.11	0.031			0.079	0.0022	0.0092	
living area	-0.022		0.75	1	0.17	0.35	0.11		-0.063	0.76	0.88		0.31	0.059	-0.08	0.055		0.76	0.18	0.0024	0.0025	0.71
lot area	0.0044	0.034	0.081	0.17	1	-0.0041	0.026	0.078	-0.0085	0.11	0.18	0.02	0.052	0.0068	0.07	-0.091		0.15	0.71	-0.013	0.0033	0.082
number of floors	-0.01	0.18	0.5	0.35	-0.0041	1	0.016	0.02	-0.27	0.46	0.53	-0.24	0.48	0.0067	-0.13	0.051		0.29	-0.01	-0.0076	0.017	0.26
waterfront present	0.012	-0.0063	0.06	0.11	0.026	0.016	1	0.4	0.019	0.08	0.072	0.085	-0.024	0.086	0.038	-0.022		0.086	0.032	0.0016	0.0014	0.26
number of views	-0.0048	0.079	0.18	0.29	0.078		0.4	1	0.053	0.25	0.16	0.29	-0.055		0.039	-0.0046		0.28	0.072		-0.0017	0.4
condition of the house	-0.027	0.027	-0.13	-0.063	-0.0085	-0.27	0.019	0.053	1	-0.15	-0.17	0.18	-0.38	-0.062	0.045	-0.003		-0.1	-0.0047	-0.0069	-0.0021	0.041
grade of the house	-0.033	0.35	0.66	0.76	0.11	0.46	0.08	0.25	-0.15	1	0.76	0.17	0.44	0.015	-0.15	0.12		0.72	0.12	0.00099	0.0049	0.67
Area of the house(excluding basement)	-0.016	0.47		0.88	0.18		0.072		-0.17	0.76	1	-0.046		0.026	-0.084	-8.8e-05		0.74	0.19	-0.0029	0.0012	
Area of the basement	-0.016	0.3	0.29	0.44	0.02	-0.24	0.085	0.29	0.18	0.17	-0.046	1	-0.14	0.075	-0.011	0.11	-0.15	0.2	0.011	0.01	0.0029	0.33
Built Year	-0.0059	0.15	0.5	0.31	0.052	0.48	-0.024	-0.055	-0.38	0.44	0.42	-0.14	1	-0.23	-0.062	-0.14	0.41	0.33	0.073	-0.0016	-0.004	0.05
Renovation Year	-0.012	0.016	0.05	0.059	0.0068	0.0067	0.086	0.1	-0.062	0.015	0.026	0.075	-0.23	1	0.018	0.029	-0.08	-0.0026	0.0059	-0.00083	0.0053	0.13
Postal Code	0.018	-0.044	-0.11	-0.08	0.07		0.038	0.039	0.045	-0.15	-0.084	-0.011	-0.062	0.018	1	-0.31		-0.11	0.077	0.011	0.012	-0.12
Lattitude	-0.023	-0.013	0.031	0.055	-0.091		-0.022	-0.0046	-0.003	0.12	-8.8e-05	0.11	-0.14	0.029	-0.31	1		0.046	-0.092	0.015	0.0072	0.3
Longitude	-0.018	0.14	0.22	0.24	0.22		-0.048	-0.08	-0.12	0.2	0.35	-0.15	0.41	-0.08	-0.099	-0.13	1	0.34	0.26	-0.01	-0.0031	0.024
living_area_renov	-0.032	0.39	0.57	0.76	0.15		0.086	0.28	-0.1	0.72	0.74	0.2	0.33	-0.0026	-0.11	0.046	0.34	1	0.19	-0.0012	-0.0057	0.58
lot_area_renov	-5e-05	0.029	0.079	0.18	0.71	-0.01	0.032	0.072	-0.0047	0.12	0.19	0.011	0.073	0.0059	0.077	-0.092		0.19	1	-0.025	-0.015	0.076
Number of schools nearby	-0.0041	0.0034	0.0022	0.0024	-0.013		0.0016	0.008	-0.0069	0.00099	-0.0029	0.01	-0.0016	-0.00083	0.011	0.015		-0.0012	-0.025	1	0.004	0.0099
Distance from the airport							0.0014					0.0029	-0.004			0.0072			-0.015	0.004	1	0.0038
- In dispose																						

print(df.describe())

min

0.000000

```
6.237575e+03
                        67.347991
                                               0.938719
                                                                    0.769934
       6.762810e+09
                     42491.000000
                                               1.000000
                                                                     0.500000
25%
                     42546.000000
                                                                    1.750000
       6.762815e+09
                                               3.000000
50%
       6.762821e+09
                     42600.000000
                                               3.000000
                                                                     2.250000
75%
       6.762826e+09
                     42662.000000
                                              4.000000
                                                                     2.500000
                     42734.000000
                                              33.000000
                                                                    8.000000
max
       6.762832e+09
                          lot area number of floors
        living area
                                                       waterfront present
count
       14620.000000
                     1.462000e+04
                                        14620.000000
                                                             14620.000000
mean
        2098.262996
                     1.509328e+04
                                            1.502360
                                                                 0.007661
std
         928.275721
                     3.791962e+04
                                            0.540239
                                                                 0.087193
min
         370.000000
                     5.200000e+02
                                            1.000000
                                                                 0.000000
25%
        1440.000000
                     5.010750e+03
                                            1.000000
                                                                 0.000000
        1930.000000
                     7.620000e+03
                                            1.500000
                                                                 0.000000
75%
                                                                 0.000000
        2570.000000
                     1.080000e+04
                                            2.000000
       13540.000000 1.074218e+06
                                                                 1.000000
                                            3.500000
max
                        condition of the house
       \quad \text{number of views} \\
                                                         Built Year
          14620.000000
                                                       14620.000000
count
                                   14620.000000
mean
              0.233105
                                       3.430506
                                                 ...
                                                        1970.926402
std
              0.766259
                                       0.664151
                                                          29.493625
```

1.000000

1900.000000

```
max
                   4.000000
                                            J. 0000000 ...
                                                             2012.000000
                                Postal Code
            Renovation Year
                                                Lattitude
                                                               Longitude
                                                            14620,000000
     count
               14620,000000
                               14620,000000
                                             14620,000000
                  90.924008
                             122033,062244
                                                52.792848
                                                             -114,404007
     mean
     std
                 416.216661
                                  19.082418
                                                 0.137522
                                                                0.141326
     min
                   0.000000
                             122003.000000
                                                 52.385900
                                                             -114.709000
     25%
                   0.000000
                              122017.000000
                                                 52.707600
                                                             -114.519000
     50%
                   0.000000
                             122032.000000
                                                 52.806400
                                                             -114.421000
     75%
                   0.000000
                             122048.000000
                                                 52.908900
                                                             -114.315000
     max
                2015.000000
                             122072.000000
                                                 53.007600
                                                             -113.505000
                                                Number of schools nearby
            living_area_renov
                                lot area renov
                 14620.000000
                                  14620.000000
     count
                                                             14620,000000
                  1996.702257
                                  12753.500068
                                                                 2.012244
     mean
                                  26058.414467
                                                                 0.817284
     std
                   691.093366
     min
                   460.000000
                                    651.000000
                                                                 1.000000
     25%
                  1490.000000
                                   5097.750000
                                                                 1.000000
     50%
                  1850.000000
                                   7620.000000
                                                                 2.000000
     75%
                  2380.000000
                                  10125.000000
                                                                 3.000000
     max
                  6110.000000
                                 560617.000000
                                                                 3.000000
            Distance from the airport
                                               Price
     count
                         14620.000000
                                       1.462000e+04
                             64.950958
                                        5.389322e+05
     mean
                             8.936008
     std
                                        3.675324e+05
     min
                             50.000000
                                        7.800000e+04
     25%
                             57.000000
                                        3.200000e+05
     50%
                             65.000000
                                        4.500000e+05
     75%
                             73.000000
                                        6.450000e+05
                             80.000000 7.700000e+06
     max
     [8 rows x 23 columns]
print(df.count())
     id
                                               14620
     Date
                                               14620
     number of bedrooms
                                               14620
     number of bathrooms
                                               14620
     living area
                                               14620
     lot area
                                               14620
     number of floors
                                               14620
     waterfront present
                                               14620
     number of views
                                               14620
     condition of the house
                                               14620
     grade of the house
                                               14620
     Area of the house(excluding basement)
                                               14620
     Area of the basement
                                               14620
                                               14620
     Built Year
     Renovation Year
                                               14620
     Postal Code
                                               14620
     Lattitude
                                               14620
     Longitude
                                               14620
     living_area_renov
                                               14620
     lot_area_renov
                                               14620
     Number of schools nearby
                                                14620
     Distance from the airport
                                               14620
                                               14620
     Price
     dtype: int64
print(df.corr())
```

```
Lattitude
                                                             0.014949
     Longitude
                                                            -0.010163
                                                            -0.001203
     living_area_renov
                                                            -0.025014
     lot area renov
                                                             1.000000
     Number of schools nearby
    Distance from the airport
                                                             0.004035
    Price
                                                             0.009890
                                            Distance from the airport
     id
                                                             -0.004542 -0.773114
    Date
                                                             0.011457 -0.027919
     number of bedrooms
                                                             -0.006157
                                                                        0.308460
     number of bathrooms
                                                             0.009206 0.531735
                                                              0.002511
                                                                        0.712169
     living area
     lot area
                                                             0.003291
                                                                        0.081992
     number of floors
                                                             0.016567
                                                                        0.262732
     waterfront present
                                                             0.001448 0.263687
     number of views
                                                             -0.001657
                                                                        0.395973
     condition of the house
                                                             -0.002136
                                                                        0.041376
     grade of the house
                                                             0.004940
                                                                        0.671814
     Area of the house(excluding basement)
                                                             0.001222
                                                                        0.615220
     Area of the basement
                                                             0.002926 0.330202
     Built Year
                                                             -0.003968
                                                                        0.050307
     Renovation Year
                                                             0.005342 0.133173
                                                             0.011528 -0.115908
     Postal Code
                                                             0.007193 0.297490
     Lattitude
     Longitude
                                                             -0.003100
                                                                        0.024414
     living_area_renov
                                                             -0.005673
                                                                        0.584924
     lot_area_renov
                                                             -0.014587
                                                                        0.075535
     Number of schools nearby
                                                              0.004035
                                                                        0.009890
     Distance from the airport
                                                              1.000000 0.003804
     Price
                                                              0.003804 1.000000
     [23 rows x 23 columns]
print(df['number of views'].value_counts())
     0
          13198
            636
     3
            351
            219
     1
     4
            216
     Name: number of views, dtype: int64
print('Mean:',df['Number of schools nearby'].mean())
print('Median:',df['Distance from the airport'].median())
print('Mode:',df['Postal Code'].mode())
     Mean: 2.0122435020519838
     Median: 65.0
     Mode: 0 122028
    Name: Postal Code, dtype: int64
print(df.isnull().sum())
     Date
     number of bedrooms
                                              0
     number of bathrooms
                                              0
     living area
                                              0
     lot area
                                              a
     number of floors
                                              0
     waterfront present
                                              0
     number of views
                                              0
     condition of the house
     grade of the house
     Area of the house(excluding basement)
     Area of the basement
     Built Year
                                              0
     Renovation Year
                                              0
     Postal Code
                                              0
     Lattitude
                                              0
     Longitude
                                              0
     living_area_renov
                                              0
     lot_area_renov
                                              0
     Number of schools nearby
     Distance from the airport
                                              0
     Price
     dtype: int64
df.dropna(inplace=True)
df.fillna(0,inplace=True)
df.interpolate(inplace=True)
```

```
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
x=df.drop(['Price','Date'],axis=1)
x.set_index(['id'],inplace=True)
y=df[['id','Price']]
```

y_pred_list = y['id'][-len(y_pred):].tolist()
y_pred_df=pd.DataFrame(y_pred_list,columns=['ID'])
y_pred_df['Predicted Price']= y_pred.round(2)

y_pred_df

x.head()

	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	grade of the house	Area of house(exclud baseme
id										
6762810145	5	2.50	3650	9050	2.0	0	4	5	10	3
6762810635	4	2.50	2920	4000	1.5	0	0	5	8	1
6762810998	5	2.75	2910	9480	1.5	0	0	3	8	2
6762812605	4	2.50	3310	42998	2.0	0	0	3	9	3
6762812919	3	2.00	2710	4500	1.5	0	0	4	8	1

y.head()

```
Price
                                 \blacksquare
      0 6762810145 2380000
      1 6762810635 1400000
      2 6762810998 1200000
      3 6762812605
                      838000
      4 6762812919
                       805000
from sklearn.model selection import train test split
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import r2_score
x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y['Price'], test\_size = 0.1, random\_state = 2)
\verb|model = GradientBoostingRegressor(n_estimators=400, \verb|max_depth=5|, \verb|min_samples_split=2|, learning_rate=0.1|)|
model.fit(x_train,y_train)
                       {\tt GradientBoostingRegressor}
     GradientBoostingRegressor(max_depth=5, n_estimators=400)
y_pred = model.predict(x_test)
model.score(x_test,y_test)
     0.9115605737002777
r2_score(y_pred,y_test)
     0.9008207044846054
y_pred
     array([497766.12740438, 244495.3776842 , 293819.40063242, ...,
             698495.60350629, 297006.00386358, 245881.76921871])
```

https://colab.research.google.com/drive/19qbihlgB-CLf7_EvkmJGC2Qq63r1jNE2?authuser=4#scrollTo=_7AySxuCReJK&printMode=true

	ID	Predicted Price						
0	6762811233	497766.13	ıl.					
1	6762811403	244495.38						
2	6762811775	293819.40						
3	6762811861	397555.35						
4	6762812009	474843.29						
1457	6762830250	1041014.57						
1458	6762830339	317512.59						
1459	6762830618	698495.60						
1460	6762830709	297006.00						
1461	6762831463	245881.77						
1462 rows × 2 columns								